个性化搜索引擎技术研究

顾立平

(国立台湾大学 图书资讯系, 台湾 台北 100671)

摘 要: 个性化搜索引擎是一种用户驱动网页排名结果的优化方式。基于本体和语义网, 用户建模可以作出准确的查询结果, 它包括: 限定搜索方式、过滤搜索结果, 以及成为搜索过程等 3 种方式。因此, 个性化搜索引擎用户模型可被视为用户驱动个性化搜索服务的模型。研究结论是整合前人研究并且提出"用户行为(用户兴趣、用户偏好、用户查询记录)-用户文档(用户行为与关键词组)-用户建模(相关性算法与排名算法)-个性化服务"的新模型, 可作为数字图书馆发展个性化搜索引擎的指引。

关键词: 信息检索; 信息搜索; 信息搜寻行为; 用户参与; 个性化数字图书馆
中图分类号: TP393.09　　文献标识码: A　　文章编号: 1672-7800(2011)04-0106-03

1 技术: 优化搜索引擎的方法

1.1 用户建模限定搜索方式

一个简单(或直接的)实现个性化搜索引擎的方式, 就是在用户搜索之前, 预设它们的用户兴趣(interest)或用户偏好(preferences)。当用户登入系统后, 系统在用户先前所指定的主题领域内, 或者文献类型内, 或者文献/网页发布时间内等, 有范围地进行检索。这是一般数字图书馆信息检索系统所采用的个性化系统模式。目前, 这种方式在个性化搜索引擎系统中的应用不多, 但是具有两个重要趋势, 值得数字图书馆参考。

(1) 整合用户兴趣的表单、用户偏好的设定以及网页排名算法, 进行个性化搜索服务。具体技术线路为: 结合经典的平面排名名单和搜索引擎, 让用户通过选择具有层次结构的文件夹标签(主题), 以交互方式查询, 在浏览过程中进行知识提取、查询优化和搜索结果个性化。这种服务模式与个性化数字图书馆相似, 但是更着重用户在浏览过程中的二次查询、根据结果进一步查询, 以及结合其它情报分析系统的辅助查询等设计。可说是个性化数字图书馆的进化版本。

(2) 从用户行为中, 建立用户文档, 将用户文档与领域本体(关键词组的关联设定)结合, 进行个性化搜索服务。具体技术线路为: 分析用户的点击记录、估计用户兴趣建立本体、利用本体替代用户当前查询的词汇。当计算用户兴趣以优化查询过程时, 需要能够有效地识别用户喜好以及为每个用户建立一个配置文件, 一旦这样的配置文件是可用的, 还需要在众多查询相匹配方案中确定用户兴趣集。因此, 这套模式的"用户行为"是指用户兴趣和用户偏好。根据这套模式, 可以发展出另一种类型的个性化数字图书馆。

如前所述, 搜索引擎和数据库检索系统的先天条件和解决问题模式不同, 目前的个性化数字图书馆系统和个性化搜索引擎也有所不同。然而, 以用户建模来限定搜索方式的个性化搜索引擎技术并不复杂, 因为它的底层技术就是在用户检索式之前, 加上系统预设的检索式, 然后进行搜索。由于搜索引擎的查询(query)多半不会要求用户输入检索公式, 而只让用户输入关键词(keyword), 所以用户仿佛感觉到这是一种个性化搜索, 事实上, 多数数字图书馆所采用的这一技术只是隐藏起部分数据库检索系统的条件式。然而, 在个性化搜索引擎当中, 其底层技术是相同的, 但是叠加技术却又千变万化, 个性化数字图书馆可予以借鉴。

1.2 用户建模过滤搜索结果

如果用户建模限定搜索结果中的用户兴趣和用户偏好交织成一张渔网, 那么用户建模过滤搜索结果中的用户兴趣和用户偏好就是一个双层漏斗。其原理是相同的, 就是把搜索结果进行删选或过滤, 前者发生在搜索之前, 后者发生在搜索之后。不过, 后者的底层技术相对来说较为复杂。目前, 这种方式在个性化搜索引擎系统中的应用较多, 具有两个重要趋势, 值得数字图书馆参考。

(1) 根据网页内容, 进行数据元(文献或网页内容的最小单位, 其概念与元数据不同, 其"元 meta"是指单位 unit 而非后设 post- 的概念)拆解与分析。具体技术线路为: 根据结构化网页记录(record)发展一项封包技术(wrapper), 包括: 以删选规则(filtering rules)过滤无关信息、以树状匹配算法(tree matching algorithm)将数据抽取提速、以频率算法检测数据元的数量和规模、以数据比对算法进行迭代和析取, 以及用合并和分割数据法来解决数据元识别的问题。这种模式可以强化元搜索引擎对大量网站数

作者简介: 顾立平(1978-), 男, 台湾台北人, 国立台湾大学图书资讯系博士后研究助理, 研究方向为决策支援系统。

据的处理速度,同时让个性化搜索引擎跨越异构资源, 在资源集成的状况下还能达到个性化服务功能。

(2) 从文献内容中抽取关键词汇, 并结合用户检索记录, 建立用户文档以进行个性化服务。具体技术线路为: 从查询结果的网页片段去识别相关查询词汇, 同时用凝聚聚类算法产生个性化查询集群, 以增强个性化搜索引擎的聚类效果; 或者, 以自组织地图算法(self-organizing map algorithm, SOM) 在用户检索后建立用户兴趣资料库, 以文本挖掘的方法来优化个性化搜索的差异性结果, 由搜索引擎提示语义相关的查询词汇的这种模式, 可以使用户可以按照反映他们信息需求的建议选择搜索词汇。

简单比较, 在用户建模限定搜索结果中, 用户预先设定了检索式, 而这个前段检索式被信息系统隐藏了起来, 如果个性化搜索得不到用户所需信息, 则要不用户承认自己原先的设定不完美, 要不用户选择全部的"用户兴趣"和取消所有的"用户偏好"(形同放弃个性化搜索) 则可获得相关信息。这种模式下的个性化数字图书馆是让找不到信息的用户"哑巴吃黄莲, 有苦说不出"。但是在用户建模过滤搜索结果中, 用户建模设定的是后段检索式, 用户在检索后, 系统自动再次检索, 并隐藏起这部分的后段检索式, 因此用户不会陷入"是否个性化"的选择, 而是进入"已经为您个性化搜索了"的过程。从某种意义上来说, 这是一种"不作恶(Don't be evil)"的作风, 也就是个性化搜索系统愿意承担用户找不到信息的责任, 而不是推卸给终端用户。

1.3 用户建模成为搜索过程

用户建模可以成为搜索引擎的渔网和漏斗, 在用户检索前后进行预先设定检索式和自动二次检索(及其相关性推荐) 的功能。用户建模也可成为魔方盒, 在用户检索中进行多重检索结果的最优化匹配。其底层技术较前两者更为复杂, 虽然建立在前两者的搜索结果和技术方法上, 但是其技术路线和前两者截然不同。它具有两个重要趋势, 是新一代个性化数字图书馆必需参考的对象。

(1) 用户建模的技术来自人工智能的应用。具体技术线路为: 基于进化理论的遗传编程 (genetic programming, GP) 学习机技术, 来优化文件在向量空间中的权重, 达到从个人查询以至不同排名结果程度上的的网页搜索排名功能; 或者, 以模糊集与模糊逻辑(fuzzy sets and fuzzy logic) 对用户满意度评分, 来优化(工作) 搜索。无论是遗传算法还是模糊逻辑, 其底层数据无非来自用户兴趣、用户偏好和用户查询。根据用户行为进行用户建模, 再转化为用户文档建立个性化服务, 已是一项发展趋势。

(2) 用户文档应用在信息检索系统和网页搜索引擎。具体技术线路为: 根据观察用户行为和行动, 动态结构化用户文档(建立用户兴趣的相关词组), 以运用在信息检索系统的延伸查询功能, 可用来改变搜索引擎排名顺序。这种技术线路的重点不是让用户建模删选和过滤搜索结果, 而是改变搜索结果, 在用户文档中的用户兴趣、用户偏好、用户查询记录和相关词组是不断改变的模式下, 用户文档参与到网页排名和文献相关性排名。

用户建模成为搜索过程的方式很多, 是未来研究个性化搜索引擎, 乃至搜索引擎的一项最主要趋势。其巨大潜力在于: 非传统意义上的用户参与(User engagement), 而它还未完全显现在搜索引擎服务中, 乃至个性化数字图书馆中。

2 应用: 优化数字图书馆的检索系统

学者用 500 个词汇查询 Google、Yahoo、Live 和 Ask 等 4 个搜索引擎, 在 42, 758 笔结果的基础上分析搜索引擎的搜索结果, 发现 Google 和 Yahoo 偏好引用自家服务(如 YouTube 和 Yahoo Answers)。数字图书馆并没有类似问题。然而, 传统的个性化数字图书馆只有 3 种个性化搜索引擎的其中一种技术, 而且较多从数据库检索系统的角度, 而非网页搜索引擎的角度来发挥个性化服务。

采用第 2 种角度, 可以丰富数字图书馆的信息组织和检索。例如, 在医疗领域中的博客(blog) 和微博客(Micro blogging) 可否算是医疗资源, 是否为数字图书馆的信息资源? 有学者研究: 病患和护士描述它们的生活, 而医生则在博客上发布保健相关信息, 这种内容差异可被搜索引擎进行排名改进, 以利用户模型搜索适当的知识来源。那么, 支持医疗团队的信息服务就需要数字图书馆的个性化搜索引擎。

电子服务(E-Service) 包括: 合作、定制、集成和适应等 4 种模式, 个性化服务的精神是个人可在协作环境下贡献、接收定制的或个性化的信息推荐、经过一个综合系统或过程, 获得及时或或时间内的支援投入。这要求数字图书馆的个性化搜索引擎能提供精确的搜索结果, 以节省终端用户在信息搜寻行为(Information seeking behavior) 所花费的时间, 好节省这段时间做其它方面工作。个性化服务从来就不止局限在个性化数字图书馆里的信息提供环节, 而是终端用户的整个工作流程中。学者研究显示: 基础科学研究员通常利用关键词在数据库或网络搜索引擎进行搜索, 而未见图书馆资源或服务整合到他们的工作中, 建议: ①图书馆资源应该可透过它们专业网站而获取; ②培养与关键行政部门的人事关系; ③集中并管理校园学术信息到机构知识库。

目前, 人们已用各种方式来建立新的数字图书馆系统。例如, 采用手动编辑用户兴趣到文本分类训练器, 个性化目录系统结合用户兴趣和分类目录, 比目录系统(categorization system, CAT) 和表单系统(list interface system, LIST) 更快、更容易发现相关信息。再如, 以本体论建立阿拉伯语和英语的产品目录检索系统(其自然语言不同需要双语本体优化搜索引擎)。又如, 根据用户文档(user profile) 建立模糊概念网络的档案检索系统, 按照用户偏好提供个性化网页和相关文件等。这些研究显示了用户模型对数字图书馆的重要性。

用户不一致的相关性判断、排名和相关性标准, 会改变个性化搜索系统的评价, 特别是对排名相似性和相关性标准随机性的测量和估计。基于这个理论, 进行"用户行为-用户文档-用户建模-个性化服务"的新模型就有其必要性。

当数字图书馆开发个性化搜索引擎时, 首先, 搜索引

擎需要能够有效地识别用户的利益, 也为个人用户建立一个配置文件; 其次, 一旦这样的配置文件是可用的, 搜索引擎需要与排名的方式相匹配的一个给定用户的利益的结果。然而, 用户不会主动地提出个人嗜好, 所以要充分利用用户的历史行为记录, 来挖掘用户行为的可能规律以及建立用户配置文档; 再次, 根据他们过去的查询记录, 即关键词语来建立可进行语义近似推理的本体论。

在这个过程中, 用户文档(User profiling) 是个性化应用的基础元素, 许多用户文档建立在用户兴趣而不是"用户不感兴趣"的内容上。透过个性化查询聚类方法, 测试用户正向负向偏好的优化策略, 可利用凝聚聚类算法(agglomerative clustering algorithm) 优化个性化查询结果。凝聚聚类算法简单来说就是一种分群算法, 首先把每个点当作一个群落, 接着透过一些距离量测方式选择要合并的点, 如此反复直到所有点都被凝聚在一起为止。这项技术的进一步延伸, 就会和创建与使用型人(creating and using personas) 结合, 成为从用户行为判断用户群组, 从用户群组和当前用户查询词汇判断信息推荐。

换句话说, 所有新一代数字图书馆系统的建立, 环绕在"用户行为-用户文档-用户建模-个性化服务"这一模型上。

3 用户驱动个性化搜索服务

在 Yahoo、Google 和 Bing 等搜索引擎逐一诞生后, 人们总是欣喜后不久就不再满足于主题内容(分类) 检索、大众化排名统计检索, 以及相近内容(聚类) 检索等服务模式。人们既想让这些技术为自己服务, 同时也想让自己的搜索引擎结果属于自己, 而非大众。这样, 就有了个性化搜索引擎的服务模式。它集成所有人的搜索过程, 包括检索式、浏览、点击和停留时间等, 作为分析用户行为模型的依据, 同时又从这种用户模型推导相关文件、排除非相关文献、进行网页或文献排名等, 再根据用户文档中的用户兴趣和用户偏好的不同, 提供个性化搜索内容。简言之, 是一种"从集成信息到分殊服务"的模式。

这种模式不难在商业活动中发现。过去, 多数 BtoC 电子商务系统的搜索引擎只能让用户搜索产品的编号、种类和价格, 而忽视可以采用代理技术去参与买家和卖家的交易。学者建议: 采用代理技术可以综合考虑价格、数量、品牌、包装、交货时间等因素, 从而优化搜索引擎, 在反复交易和检索过程中, 达到个性化推荐最适合的产品给当前用户。这种技术可以为数字图书馆参考。

本文系统地梳理优化搜索引擎的 3 种用户建模方式: 限定搜索方式、过滤搜索结果以及成为搜索过程等。对优化数字图书馆检索系统, 提出"用户行为(用户兴趣、用户偏好、用户查询记录) - 用户文档(用户行为与关键词组) - 用户建模(相关性算法与排名算法) - 个性化服务"模型。

其中, 用户建模成为搜索过程还有许多研究空间。如果用户意向被更好地运用, 则能够一般化文本片段抽取(text snippet extraction) , 比方使用统计语言模型捕获文档和用户意向的共性。而采用类似网页排名(PageRank) 的实例算法(InstanceRank) 能减少实例集的大小, 从学习库中选择最有代表性的实例。在扩展个性化搜索引擎的未来研究方面, 有研究显示社会媒体网站构成了相当一部分的搜索结果。这说明了社会媒体作为个性化搜索引擎的一部分。此外, 透过连续主成分分析(continuous principal component analysis, CPCA) 进行傅里叶级数(Fourier series) 计算, 以求得三维模型的平移(translation) 、旋转(rotation) 、翻转(flipping) 和大小(scale) , 能够让搜索引擎用形状相似性搜索三维模型数据库。这说明了虚拟社会可作为个性化搜索引擎的一部分。这些研究方向值得后续追踪和探索。

参考文献:

[1] PAGE L, BRIN S, MOTWANI R, WINOGRAD T. The page rank citation ranking: bringing order to the Web (1999) . [ED/OL] [2010-10-27] http: / / ilpubs. stanford. edu: 8090/ 422/ .

[2] KIM H, CHAN PK. Personalized search results with user interest hierarchies learnt from bookmarks [J] . Advances in Web mining and Web usage analysis, 2006.

[3] JIANG X, TAN AH. Learning and inferencing in user ontology for personalized Semantic Web search[J]. Information sciences, 2009(16) .

[4] KE YP, DENG L, NG W, LEE DL. Web dynamics and their ramifications for the development of Web search engines[J] . Computer networks, 2006(10) .

[5] CAMBAZOGLU BB, KARACA E, KUCUKYILMAZ T, TURK A, AYKANAT C. Architecture of a grid-enabled Web search engine[J] . Information processing and management, 2007(3) .

[6] STEGERS R, FEKKES P, STUCKENSCHMIDT H. MusiDB- A personalized search engine for music[J] . Journal of Web semantics, 2006(4) .

[7] BAR-LLAN J, KEENOY K, YAARI E, LEVENE M. User rankings of search engine results[J] . Journal of the American society for information science and technology, 2007(9) .

[8] FERRAGINA P, GULLI A. A personalized search engine based on Web-snippet hierarchical clustering[J] . Software-practice & Experience, 2008(2) .

[9] STAMOU S, KOZANIDIS L, TZEKOU P, ZOTOS N. Ontology-driven personalized query refinement[J] . Journal of Web engineering, 2009(2) .

[10] HONG JL, SIEW EG, EGERTON S. Information extraction for search engines using fast heuristic techniques [J] . DATA & KNOWLEDGE ENGINEERING, 2010(2) .

[11] LEUNG KWT, NG W, LEE DL. Personalized concept-based clustering of search engine queries[J] . IEEE transactions on knowledge and data engineering, 2008(11) .

[12] HUNG CL, CHI YL, CHEN TY. attentive self-organizing neural model for text mining[J] . Expert systems with applications, 2009(3) .

[13] FAN WG, PATHAK P, WALLACE L. Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search[J] . Decision support systems, 2006(3) .

[14] GURSKY P, HORVATH T, JIRASEK J, KRAJCI S, NOVOTNY R, PRIBOLOVA J, VANEKOVA V, VOJTAS P. User preference Web search - experiments with a system connecting Web and user[J] . Computing and informatics, 2009(4) .

[15] KUMAR H, KANG S. Exclusively Your's: Dynamic Individuate Search by Extending User Profile[J] . New generation computing, 2010(1) .

(责任编辑: 周晓辉)